

Summary

Summary

Methodological News features articles and developments relating to work done within the Methodology division. This issue contains five articles:

- Investigation into application-based positioning data for COVID-19
- Exploring the use of machine learning for anomaly detection and editing
- Development of a monthly partial turnover indicator
- Text analytics
- Time series clustering

Investigation into application-based positioning data for COVID-19

Investigation into application-based positioning data for COVID-19

As part of its COVID-19 response, the ABS investigated application-based geospatial positioning data from mobile phones as a source of information on social and economic changes during the coronavirus crisis.

Many ad-supported mobile applications ('apps') use Global Positioning System (GPS) data to serve location-specific advertising, if users have given explicit permission for the apps to access that positioning data. In recent years researchers have started using app-sourced location data to explore social issues including economic and epidemiological aspects of the COVID-19 crisis.

In 2020 the ABS purchased two samples of app-based location data from a commercial provider, covering the state of Victoria for January-April 2020. These samples had been aggregated to Statistical Area Level 1 (SA1) and whole-of-state level. No information about individuals was obtained, nor any information that could be used to identify individuals. The data is not related in any way to the Australian Government's COVIDSafe app.

The first sample describes 'close contact' events where two mobile devices are observed within five minutes and five metres of one another (five metres being the limit of GPS precision). One version of the aggregated data assigns contacts to the regions where they occur ('event-location'); the other assigns them to the 'common evening locations' of the devices involved.

Comparison to publicly available sources of data (Apple and Google mobility data, City of Melbourne pedestrian counters) gave broadly consistent results, with all sources showing a sharp drop-off in activity from mid-March 2020, but the 'close contacts' data has two major advantages.

First, it allows for state- or nation-wide coverage at fine levels of geography, something not available from the public sources.

Second, the ability to view both 'event-location' and 'common evening location' data offers insight into patterns of activity. For instance, contacts may increase within a residential suburb during lockdown, because residents who previously commuted elsewhere during the day are now spending more time in their home suburb. Examining the "common evening location" view helps understand this pattern: residents are making fewer contacts overall and these contacts have moved from the city to the suburb.

The second sample gives 'foot traffic' data: daily counts of devices observed at various locations of interest. These locations are primarily retail stores for major brands, but also include some other location types such as universities, Centrelink offices, and national parks. Comparison with other sources available to ABS showed that for many categories of retail business, foot traffic tracked closely with business income.

The ABS is now considering further applications for this type of data.

For further information, please contact Geoffrey Brent at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Exploring the use of machine learning for anomaly detection and editing

Exploring the use of machine learning for anomaly detection and editing

The ABS is assessing the feasibility of using machine learning to identify and correct anomalous data. This work aims to understand how machine learning can help to prioritise and reduce manual effort, identify incorrect data, find new edit rules, propose edit value(s) with a measure of confidence, and address issues that traditional approaches struggle with.

We are currently focussed on business administrative data, particularly big data (in terms of number of records and data items) with a time-series aspect.

As part of this work we are addressing key aspects such as understanding the relative advantages and limitations of different methods, including traditional approaches, and the use of several techniques in combination. We will also explore:

- Methods to adequately demonstrate and explain how a machine learning algorithm came to a solution.
- How to develop edit rules for new or evolving datasets.
- How to update models and algorithms over time and in response to real-world changes.

We are assessing a selection of machine learning and traditional approaches, ranging from simple techniques to more-complex methods suited to the time-series nature of some datasets. The aim is to develop a toolbox of techniques suited to datasets with different characteristics.

Initial work has focussed on data where there is limited understanding of the behaviour patterns of correct vs wrong data, where a key challenge is to develop a labelled dataset that the methods can be assessed against. Unsupervised approaches such as Local Outlier Factor can assist to identify anomalies, however human expertise is needed to help differentiate wrong data from unusual data.

The ABS is engaging with other National Statistical Organisations who are undertaking similar work in order to share best practice.

For further information, please contact Jenny Pocknee at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Development of a monthly partial turnover indicator

Development of a monthly partial turnover indicator

Businesses provide Business Activity Statements (BAS) to the Australian Taxation Office (ATO) as part of their tax obligations. BAS contains useful information including business turnover, wages and capital expenditure. The ATO provides BAS data to the ABS to be used in statistical production processes. These include the formation of homogenous strata in sample designs and forming benchmark totals for survey estimation.

The ABS is looking to leverage off already available administrative data sources to produce more timely insights into the Australian economy. One such area of recent research has been the production of a monthly business output indicator based on BAS turnover. That is, instead of using BAS turnover just to aid in statistical production, we are exploring how to use these data as the primary source of a new monthly economic partial indicator (PI). This indicator will utilise turnover data from businesses which submit BAS turnover to the ATO on a monthly basis.

The ABS has investigated the challenges in producing a monthly PI. One of the key issues is addressing missingness in the monthly data from:

1. Late responses - businesses submit their BAS form too late to be included in the PI for that particular month.
2. Deaths - businesses don't submit a BAS form as they are no longer operating.
3. Changes in submission frequency - this typically involves a business switching between monthly and quarterly

submission of BAS data.

It is challenging to address the missingness as it is not known at the time of non-submission which of the above is the cause. Historical imputation is used to address late response (Type 1 above), where the late submitters' data is imputed using the most recent reported BAS data for the unit. A factor is applied to account for changes in the PI between months. This is derived from the change between continuing responding units in the current month and the most recent reported month.

To account for deaths (Type 2), an additional factor is applied to the historical imputation. This factor estimates the likelihood a business is still alive, and hence should have reported BAS. The likelihood is derived from historical data by observing the frequency of non-submitters subsequently submitting in another month. Probabilities will vary depending on the number of consecutive months of non-submission (note some businesses may not submit for several months due to tardy BAS reporting).

Changes in submission frequency (Type 3) are addressed through a coverage adjustment. This is still being investigated, but the current approach is to apply a weight representing the contribution of quarterly submitters, so as to increase the level of the PI to reflect both quarterly and monthly submitters. This weight is derived from the most recent quarter of BAS submission. It is hoped that this will protect against monthly changes in the PI being impacted by changes in submission frequency.

Quality assessment of the PI has been undertaken throughout its development, where comparable ABS survey statistics have served as a reference point. These assessments have been promising, and future refinement of the PI methodology should further improve its performance.

For further information, please contact Tom Davidson at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Text analytics

Text analytics

As part of its research program in text analytics – the automatic extraction of information content from textual data sources – the ABS is investigating the potential use of Named Entity Recognition (NER) in statistical production. NER is a capability in the field of Natural Language Processing that detects references to entities in text and classifies them into predefined categories that have similar attributes, such as a person, business, location, product or event. A named entity is one that is identified by a distinctive word or phrase. For example, a person or business may be identified by a name, a location by address, and an event by date.

There have been many approaches to NER systems in the past, including rules-based identification and categorisation, unsupervised methods such as clustering, and feature-based supervised learning. Recently, the use of neural networks has delivered state-of-the-art results in entity extraction, and this approach is being investigated for use in ABS applications.

The two current use cases for NER in the ABS involve identifying and extracting descriptive information about:

- Consumer products (such as type, model, version, size) contained within retail scanner transactions records.
- Job vacancies (such as position title, required skills, location) contained in web-scraped online job advertisements.

The job vacancy use case has been chosen as the initial focus area for investigation. Following a survey of the literature in December 2020, an experiment was conducted using web-scraped Australian job advertisements data from Seek, which are publicly available on the Kaggle data science portal. The motivation for this work is to create entity-level data about jobs that can be utilised in the ABS Labour Market Analysis Project (LMAP) to deliver new insights on the impact of COVID-19 on employment and labour demand. This project also uses GLIDE – an advanced knowledge discovery system under development by the ABS.

The Seek dataset required manual annotation to create training data for the machine learning process. The Python package spaCy was then used to generate a basic NER language model to extract key job characteristics. The next stage of research aims to develop a machine-interpretable concept model of job classes, and to iteratively improve the performance of the language models for use in the LMAP.

For further information, please contact Phil Newbold at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

Time series clustering

Time series clustering

The ABS is analysing the impact of the COVID-19 pandemic on the Australian economy using new methods, technologies and sources of data. One new method under investigation is the application of specialised clustering techniques to find patterns across large numbers of different time series. A time series is simply a set of measurements made sequentially in time. Clustering is the task of grouping objects in such a way that those in the same group (called a cluster) are more similar to each other (according to some objective measure) than to those in other groups.

Clustering different time series by their structure exposes similar features that may be caused by common events or behaviours in the underlying system of interest. This approach is more efficient than manually inspecting each individual time series, and it also avoids biases caused by subjective interpretation. For example, given data about personal mobility by country or state, can regions be identified that have been similarly affected by the lockdown measures imposed by governments, based on the types of business/industry or the demographics of residents in these areas? Or given thousands of time series of merchandise exports and imports data, can categories be found that give a better understanding of the factors influencing global trade patterns?

There are many techniques for time series clustering. However, the main issues to be considered for any time-series clustering approach are: how to represent the time series in a compact way that still preserves its essential characteristics; how to measure the similarity between time series under conditions that may cause distortions; and what algorithm to use for the clustering process.

To explore some of these issues, an undergraduate student from Monash University undertook a research project during the 2020-21 summer break as part of the ABS Vacation Student Program. The objectives were to evaluate the effectiveness, performance and computational demands of alternative approaches for clustering time series, using publicly available Apple and Google personal mobility data and open source R software packages in an AWS cloud computing environment. The work covered different representation schemes (wavelet transform, ARIMA), similarity metrics (Euclidean, dynamic time warping) and clustering algorithms (partitioning, hierarchical decomposition).

Some key findings from the project were that:

- The large volume of data involved makes it essential to reduce the dimensionality of the problem – by appropriately summarising the time series – for large-scale computation.
- The results of clustering depend critically on how the similarity of clusters is quantified – whether the focus is on trends, seasonal/weekly patterns, or volatility. It is therefore not possible to identify one clear ‘best’ method, as the most suitable technique will depend on the goals of the analysis.

Further work is planned over the course of 2021 to improve and deploy basic methods and software for a wide range of analyses, starting with an investigation of merchandise trade data in April 2021. Research will then extend into the application of machine learning to time series classification and event detection.

For further information, please contact Ric Clarke at methodology@abs.gov.au.

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

How to contact us and email subscriber list

To provide comments or feedback, or to be added to or removed from our electronic mailing list, please email methodology@abs.gov.au or contact:

Methodological News Editor
Methodology Division
Australian Bureau of Statistics
Locked Bag No. 10
Belconnen ACT 2617

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.

About this Release

Methodological News features articles and developments relating to work done within the Methodology division.

© Commonwealth of Australia

All data and other material produced by the Australian Bureau of Statistics (ABS) constitutes Commonwealth copyright administered by the ABS. The ABS reserves the right to set out the terms and conditions for the use of such material. Unless otherwise noted, all material on this website – except the ABS logo, the Commonwealth Coat of Arms, and any material protected by a trade mark – is licensed under a Creative Commons Attribution 2.5 Australia licence